DOCUMENT RESUME

ED 347 809 FL 020 173

AUTHOR Fox, Janna; And Others
TITLE Psychometric Properties of the CAEL Assessment: An Examination of the Dependability/Reliability of Placement Decisions.
PUB DATE 22 Jun 92
NOTE 36p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Criterion Referenced Tests; *English for Academic Purposes; Foreign Countries; Higher Education; *Language Skills; *Language Tests; Limited English Speaking; *Psychometrics; Second Language Instruction; *Student Placement; Test Reliability
IDENTIFIERS *Carleton Academic English Language Assessment; Carleton University ON

ABSTRACT

The Carleton Academic English Language (CAEL) Assessment is a criterion-referenced test developed at Carleton University (Canada). It combines logic of performance testing with ethics of gradual admission, i.e., that limited-English-speaking students should be allowed to begin limited academic study in English, based on language proficiency and with the support of additional language instruction. The test was developed from the model of the Ontario Test of English as a Second Language, but provides multiple, parallel versions for test security. The test is administered in two stages: (1) student registration by trained interviewers, with an informal sample of student writing taken; and (2) a formal, written test based on a simulation and set of integrated activities on a single academic theme, drawn from a first-year undergraduate course. Using results, it is possible to construct a profile of language use for academic purposes. Each language skill (speaking, listening, reading, and writing) is evaluated differently but standardized to a band score. The four scores provide data for overall language assessment and placement decisions. The test's reliability for placement decisions was assessed through teacher perceptions of placement appropriateness, advisor feedback, and grade trends. It is proposed that the CAEL may provide a model for other institutions. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED347809

PSYCHOMETRIC PROPERTIES OF THE CAEL ASSESSMENT: AN EXAMINATION OF THE DEPENDABILITY/RELIABILITY OF PLACEMENT DECISIONS

Janna Fox, Centre for Applied Language Studies
Carleton University

Bruno D. Zumbo, Measurement and Evaluation
University of Ottawa

Timothy Pychyl, Department of Psychology
Carleton University

June 22, 1992

FL 020 173

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Janna Fox

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

BEST COPY AVAILABLE

> . . .validation [is] a general process that consists of the marshaling of evidence to support a given interpretation or use, a process that is based on logical, empirical, and ethical considerations (Bachman:1990)

THE CARLETON ACADEMIC ENGLISH LANGUAGE (CAEL) Assessment is a criterion referenced test which is part of a collaborative testing system developed at Carleton University, Ottawa, Canada. It combines the logic of performance testing with the ethics of gradual admission. It is logical that students should be tested on language in use for academic purposes if that is precisely the way in which they intend to use language. It is ethical to recognize degrees of proficiency, and allow students to begin study in their academic field on a limited basis, based on their degree of proficiency in English, with the support of one or more English for Academic Purposes (EAP) courses.

This paper presents an overview of test development, examines empirical data relating to the reliability and dependability of CAEL placement decisions, and suggests that the CAEL testing system may provides a model for other universities, colleges, and institutions who wish to generate positive washback effects from language testing for purposes of admission to academic programmes.

ORIGIN OF THE COLLABORATIVE MODEL OF IN-HOUSE TESTING

In developing the CAEL we responded to concerns from EAP professors, students seeking admission to the University, and graduate and undergraduate faculty.

EAP professors were concerned about the negative washback effects of external tests on their programmes. Students who wished to gain entrance to the University needed to perfect the test-taking strategies that would allow them to pass TOEFL or MELAB, and exerted considerable pressure on EAP teachers to help them "train for the test". This kind of training, however, was inconsistent with what EAP professors knew was in the long term interest of their students -- the academic skills required for learning.

In addition, students who were unable to gain admission to undergraduate programmes because of TOEFL scores that did not meet

the entrance requirements, continuously made appeals that they be allowed to begin their programme on a limited basis. They argued that although they did not have the level of proficiency in English required for full admission, they did have a degree of proficiency, and therefore should be allowed to start their programme on a limited basis.

Undergraduate and graduate professors in the academy as a whole were also concerned about the level of proficiency of some of the students who were entering their programmes. The initial response of the faculty was to raise TOEFL entrance requirements to 580 for undergraduate students. Some graduate departments began to unofficially apply ever-higher TOEFL filters (in some cases over 600). It became increasingly obvious that simply raising TOEFL requirements would not address the problem. For example, graduate departments which had raised TOEFL requirements, continued to receive students who were funded by the University to act as Graduate Teaching Assistants, but who were unable to communicate with their students or write papers with the fluency required.

Thus, in this context, there was a general consensus at the University that an alternative testing system should be developed, in-house, which would respond directly to the concerns above. By motion of the Senate in 1987 new University regulations were approved to fund and implement the University's new CAEL testing system. TOEFL and MELAB continued to be accepted by Carleton for admission purposes, but whereas TOEFL scores were interpreted to mean that a student was either ready for full-time study or not, the CAEL Assessment was intended to provide a formula for gradual admission into full-time study. In this way, the CAEL Assessment was designed to:

1) identify students who are able to meet the linguistic demands of full-time study, or
2) place students in English for Academic Purposes (EAP) credit courses while they take one or more courses in their field of study, or
3) identify students who require full-time ESL/EAP in preparation for later studies.

THE PROCESS OF TEST DEVELOPMENT: ROLE DELINEATION

Initially, the testing committee decided that the Ontario Test of English as a Second Language, produced by the Province of Ontario, might provide the answer. It was a carefully constructed, extensively examined 'performance' test, and a number of the contributing members to OTESL worked at the University. There were, however, only two versions of OTESL, and it became obvious that although OTESL was an excellent 'performance' test, it was essential that multiple, parallel versions be available for test security if testing was to be a permanent and on-going activity. More importantly, having involved faculty from all disciplines, the test committee concluded that the development of an in-house test/tests would generate continued cooperation. One of the hallmarks of the CAEL model is its collaborative, generative potential for positive washback.

Thus, early in 1988 a group of EAP professors in the Centre for Applied Language Studies began to work with professors in the faculties of Science, Engineering, Social Science and Arts collecting information about actual 'performance' requirements in introductory, first-year classes. A number of language activities were consistent across all faculties. For example, students need to be able to take information from lectures. Although students exhibit a multitude of different note-taking strategies during lectures, when asked by professors to fill in diagrams, flow charts, or label items on a map, they exhibit very similar responses. Such consistent responses also emerge when professors provide their own definitions of key terms, when professors signal important information with discourse markers such as, "ok ... and these are really the three key points ..." or "so the main thing I want you to remember is ..." etc. It was possible, in this way to identify and catalogue consistent lecture behaviour of students, across faculties at the University, and to record lectures in progress. Students were asked to bring their notes to follow-up, small group discussion sessions and photocopies were made of actual student notes and analyzed to identify consistent student responses to lectures.

All of the first-year lectures which were observed during the test development process either introduced, or extended information which students were to read or had read. In addition, all of the lectures occurred during September and October, at the beginning of the academic year, when professors were often lecturing on information which they assumed their students knew nothing about. At the time the lectures were being recorded, lists of articles, text book readings, etc. were provided by the professors along with possible test questions which the professors would ask as a follow-up to the lecture and reading. Finally, all of the tests, lab and essay assignments required of students in the introductory courses during the year were also examined.

A review of test, lab and essay responses across the University revealed that only two departments, Psychology and Economics, used multiple choice items -- and then only in first year. All other departments consistently used short or long essay responses -- even the department of Mathematics used some extensive writing as students were asked with increasing frequency to describe the process they used in solving problems.

As a result of this role delineation, and using OTESL as a model, several versions of the CAEL were developed for pilot testing by September 1988. They consisted of actual lecture segments taken from introductory lectures at the first-year level from the faculties of science (biology) and social science (geography). The topics of the lectures were general (food, weather), but the information provided by the professor was specific to that professor's particular emphasis in the course. In all cases the professor's agenda was to introduce to his/her students information about a general topic, which the professor assumed the students knew nothing about. Students were asked to fill in diagrams, flow charts etc. which were actually part of the source lecture. They were asked to read about the same topic, and answer questions about the reading in much the same way as the students in the first-year class had been asked to do. The reading and the lecture provided a "scaffold" for the students' written response, which was the natural outcome (as it had been in the

source classes) of the information provided.

These versions of the CAEL were pilot tested with students in a local high school, students in ESL pre-University level classes, in credit ESL classes, and in the context of Introductory Psychology classes with fully admitted first-year students, who reflected the University as a whole, both in their range of ability and in their faculty of study.

CAEL ASSESSMENT FORMAT

CAEL is administered in two stages as described below [see Figure 1 on the next page for summary]:

First, all students must register in person in stage one of the assessment. During the registration process, students are interviewed by trained interviewers who holistically assess their level of oral language and provide them with information about CAEL, including a practice version of the test. This holistic impression depends on evaluation by a trained interviewer.

During registration a sample of the students' writing is also collected. Students write informal, personal responses to questions which, for example, ask them to comment on their strengths and weaknesses in English, or their plans with regard to academic study. This writing sample, generated in an informal atmosphere where testees may talk to friends, and are allowed to respond without the usual test contraints such as carefully imposed time limits,etc. provides an additional source of information about the testee's writing in a personal mode.

The writing sample is used when placement decisions are difficult because, for example, a student fails to provide a "sufficient" response (i.e. one which is sufficiently long, one which is written in the student's own words as opposed to being copied directly from reading booklets, etc.). Writing research indicates that students express themselves with greater fluency, confidence and control in the personal/expressive mode [Britton: 1987; Freedman and Pringle: 1986] as opposed to the transactional mode (argumentative or expository) required by most academic study.

There is an interval of at least 24 hours between the CAEL

interview and the administration of the written CAEL in stage 2. Students are encouraged to contact resource teachers for additional information about the CAEL if questions arise as they are working through the sample test.

The second stage is the written portion of the CAEL: an integrated set of language activities which revolve around a single academic theme -- taken directly from a first-year, introductory level course at Carleton University in Social Science, Arts, Science or Engineering. Students experience a mini-lesson on general topics such as food, health, weather etc. where a professor is introducing a particular bit of "new" information to a group of students whom he/she anticipates know nothing about the subject (as is often the case in first-year, introductory classes). To complete the assessment, students read two or more short articles about the topic (with accompanying comprehension tasks), they hear a tape of the actual university lecture on the topic (while responding to comprehension items in the same way that students' in the original class from which the sample was taken were asked to respond (i.e. by taking notes, writing definitions, filling in flow charts, diagrams etc.), and then with their readings in hand and their notes from the lecture, they write a short essay.

It is important to note that the readings and lecture supplement one another and provide the basis for the writing task. The writing prompts are actual tasks set by professors in their introductory courses and take a number of forms; for example, compare and contrast, or present an argument.

Based on these three activities in Stage 2 -- reading, listening and writing -- and the oral interview in Stage 1, it is possible to construct a profile of the students' use of language for academic purposes. The profile forms the basis of the assessment and placement decisions. As a result, if students have particular language needs, these can be addressed as identified by their profiles, and thus, the formative information provided by the CAEL is useful to both teachers and students alike.

Figure 1

STAGE 1: Prior to the formal, written CAEL, helps to provid a more complete picture of overall language proficiency, an information which assists in final placement decisions.

Information obtained about the testee:

1. one-on-one informal interview [Holistic evaluation of oral language]
2. informal writing sample [Information about the student's writing in a personal mode]
3. background information [academic; language; status of registration; year of study etc.]
4. verification of personal identification for test security

Information provided to the testee:

1. an overview of the CAEL assessment, a description of the test (including a sample test)
2. information about the Univesity's system of admission and admission options

STAGE 2: The formal, written portion of CAEL: simulation (1½2 - 2 hours in length) including lecture, readings and writing task on the same topic.

Information obtained about the testee:

1. written responses to a lecture [listening score]
2. written responses to questions about readings extending information in the lecture [reading score]
3. written response to a writing task which draws on information presented in the lecture and the reading [writing score]

Information provided to the testee:

1. profile of language proficiency in using English for academic purposes
2. placement in full-time ESL, part-time or full-time study with EAP credit courses required, or in full-time regular study without further language requirements

SCORING

Each of the four language skills -- speaking, listening, reading and writing -- are evaluated in a different manner. Despite these differences in method, each assessment is standardized to a band score. The scoring applied in each of these contexts is described below:

Oral proficiency is assessed globally or holistically by the interviewer. Once rapport has been established, the interviewer speaks with the student about a variety of pre-determined topics. The criteria used to assess the student's oral proficiency include (but are not limited to): the accuracy of the student's discourse, his/her need for clarification or repetition, ability to present ideas coherently and continuously, ability to initiate and follow topic change, and general fluency. This assessment is quantified directly as a standardized band score ranging from a low of 10 to a high of 90 (see Figures 2 and 3).

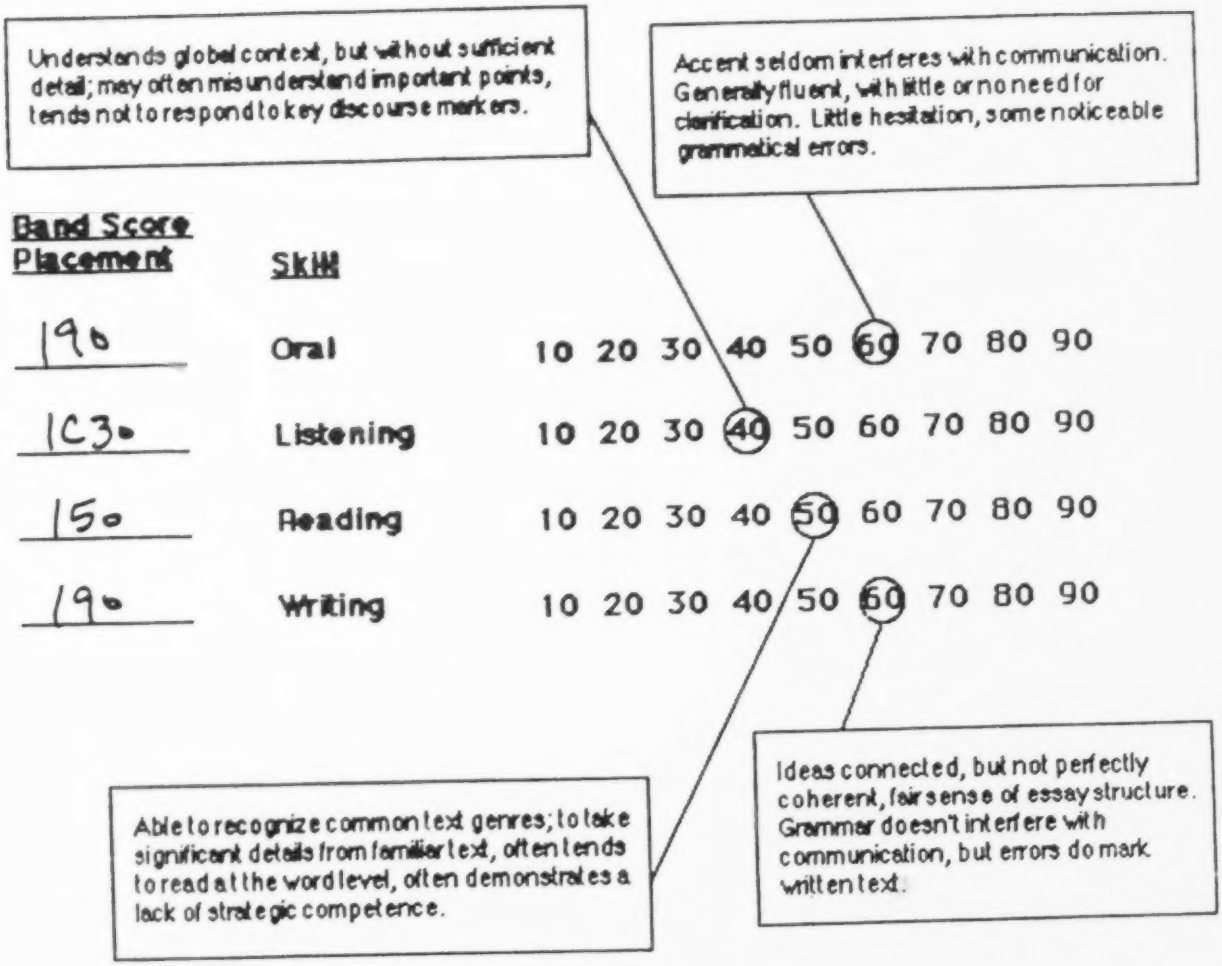
In order to assess **listening proficiency**, students are required to respond to a variety of items while listening to a tape recorded university lecture. Students hear the lecture only once, and are required to respond in much the same way as the students in the actual lecture were required to respond, by filling in information on flow charts, maps, tables; defining key concepts; summarizing the key points as identified by the professors. The responses are scored for accuracy, and the raw score is transformed to a standardized band score (see Figures 2 and 3).

Reading proficiency is assessed by a structured set of questions about the assigned readings. The questions are generally divided into two sections; questions about information that can be taken from the article by skimming and scanning for either main ideas or specific details. These questions are scored for accuracy, and the raw score is transformed to a band score for reading (see Figures 2 and 3).

Writing proficiency is based on an evaluation of the student's response to a short essay question. This essay question requires the student to draw on the information from the lecture and the readings, and to integrate this information into a coherent, structured argument. The assigned questions have been sampled from the types of questions that are required by university professors in assignments and exams. This short essay is evaluated separately by three examiners for organization, content and langauge use. Any differences in assessment are discussed, and a final assessment is made as a consensus among the three examiners. This raw score is transformed into a band score for writing (see Figures 2 and 3).

Figure 2

Example of band score profile - Credit Course Level

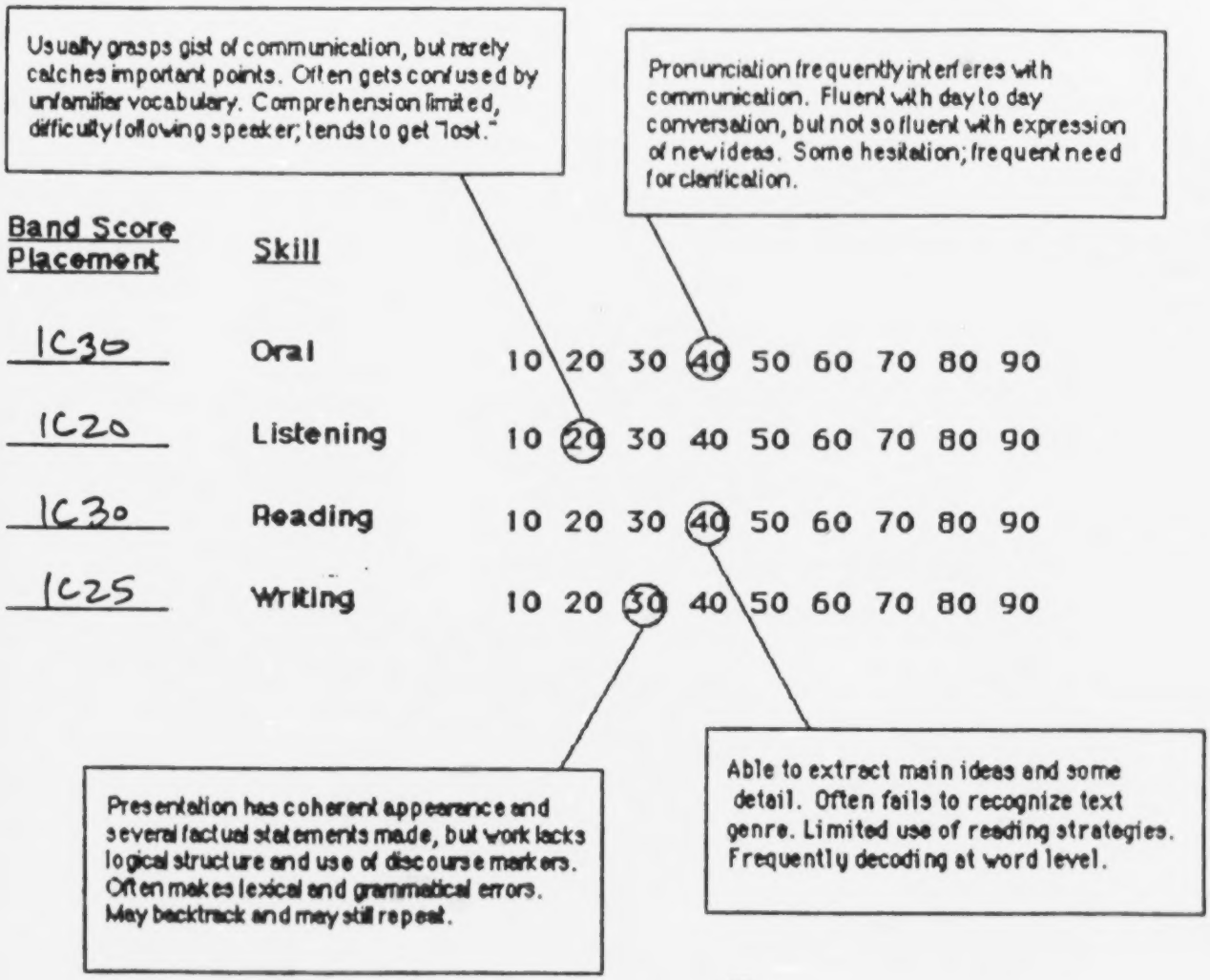

Understands global context, but without sufficient detail; may often misunderstand important points, tends not to respond to key discourse markers.
Accent seldom interferes with communication. Generally fluent, with little or no need for clarification. Little hesitation, some noticeable grammatical errors.
Band Score Placement
Skill
Oral
Listening
Reading
Writing
10 20 30 40 50 60 70 80 90
Able to recognize common text genres; to take significant details from familiar text, often tends to read at the word level, often demonstrates a lack of strategic competence.
Ideas connected, but not perfectly coherent, fair sense of essay structure. Grammar doesn't interfere with communication, but errors do mark written text.

Overall Placement: 150 / 190

BEST COPY AVAILABLE

Figure 3 (continued)

Example of band score profile - Intensive Course Level


Usually grasps gist of communication, but rarely catches important points. Often gets confused by unfamiliar vocabulary. Comprehension limited, difficulty following speaker; tends to get "lost."
Pronunciation frequently interferes with communication. Fluent with day to day conversation, but not so fluent with expression of new ideas. Some hesitation; frequent need for clarification.
Band Score Placement
Skill
IC30
Oral
10 20 30 40 50 60 70 80 90
IC20
Listening
10 20 30 40 50 60 70 80 90
IC30
Reading
10 20 30 40 50 60 70 80 90
IC25
Writing
10 20 30 40 50 60 70 80 90
Presentation has coherent appearance and several factual statements made, but work lacks logical structure and use of discourse markers. Often makes lexical and grammatical errors. May backtrack and may still repeat.
Able to extract main ideas and some detail. Often fails to recognize text genre. Limited use of reading strategies. Frequently decoding at word level.

Overall Placement: IC30

INTERPRETATION OF CAEL SCORES AND PROGRAMME PLACEMENT

The four band scores described above form the basis of the overall assessment of language competence and program placement. These four band scores create a profile of the student's language proficiency, and this profile is used to make placement decisions. The profile of language proficiency is essential to the placement process as it provides an assessment of the student's relative strengths and weaknesses linguistically.

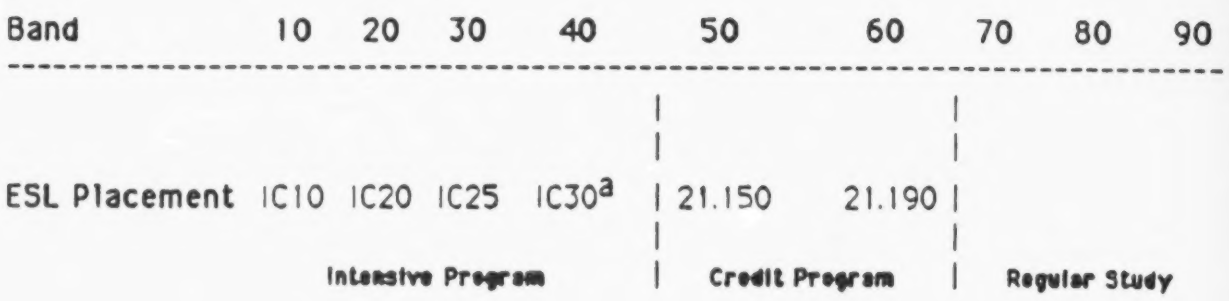
The key band score in placement decisions is writing. The performance on the writing task is the direct result of information available to the student from the lecture and the readings. Information about the student's performance in listening and reading is recorded separately on the student's profile principally for formative purposes. For purposes of placement, however, the two sections are considered together.

Because tasks are sampled from actual university courses, lectures and readings vary in relative difficulty. When readings are difficult, professors often supplement them with lectures to support and extend the information. The reverse situation also exists. Lectures which deal with complex or difficult ideas are often explained in simpler terms by supporting readings. Thus, the CAEL listening and reading sections are not seen as entirely independent or "pure" measures of listening and reading, but rather as tasks which support or scaffold the writing task.

Figures 2 and 3 above illustrate the relationship between students' band scores and placement. In Figure 4 we present an overview of student placement based on general band scores.

Figure 4

Course Placement in Relation to Band Scores



Legend for Course Numbers
IC = Intensive Course (Non-credit Program)
21.150 = Intermediate EAP
21.190 = Advanced EAP (B- required in this course to meet language requirements)

[a]Graduate students are not generally placed in IC30. Graduate students whose profile place them at the IC30 level are placed in a special credit course - 21.130. Note: 21.130 is not available to VISA students.

CAEL OVERVIEW: ISSUES OF RELIABILITY PLACEMENT DECISIONS

The CAEL provides a practical, common-sense approach to the problems associated with the assessment of language in use for academic purposes. The CAEL attempts to simulate in a test situation, the academic context which students are seeking to enter: to provide students with the kinds of tasks that they will actually face when they begin to study at the University.

In doing this it demonstrates some very important aspects of validity. The language tasks for the assessment have been carefully sampled from the academic community, and with its direct assistance to provide a sound basis for validity. As the tasks themselves are real samples of the kind of language tasks that are demanded in the academic community, the test has obvious validity for the students as well. According to surveys of 100 testees during the 1991-92 academic year, 95% responded positively to CAEL with regard to "fairness as an evaluation of English language proficiency". Five (5%) indicated they would have performed at a higher level if the topic had been more directly related to their field of study. Although test-retest data do not suggest that topic affects performance, in order to accommodate the 5% who preferred to write on topics from their discipline of study, students are now informed in advance of the Faculty of origin of test topics. [See test-retest results below.]

The test also has validity for EAP teachers, as the test asesses those elements of EAP which make up day-to-day clasroom experience, and the EAP teachers themselves work on test development along with the broader University faculty.

Statistics collected regarding student placements in EAP classes consistently indicate that teachers find the placements accurate. Subsequent to every testing period, and after an initial two to three weeks of class, teachers are asked to identify any students whom they feel "may have been misplaced". Over the past academic year 945 students were assessed by the CAEL. Teachers have identified 34 of the 478 (7%) who were placed in and actually registered for low, intermediate and advanced EAP credit courses as inappropriately placed.

Of the 20 who were identified by teachers as "performing below the required level", 18 did not attend classes regularly, had family or financial problems which prevented them from focussing on the course, or simply failed to work up to the levels expected, although their level of language ability seemed to be appropriate to the demands of the course. Sixteen (3%) were identified by teachers as "having the potential to perform well at a higher level of placement". Teachers identified only 5 (1%) students who should "definitely have been placed at a higher level".

In summary, based on data collected from the 1991-92 academic year 7 students (1%) of the 478 registered in intensive and EAP courses were "misplaced by CAEL" according to their EAP teachers. These students, and others identified as a control group, are part of a research study being undertaken by the Centre for Applied Language Studies. Data from this study will be available in 1994.

Feedback from undergraduate and graduate supervisors at Carleton also indicates general satisfaction with the CAEL screening. A recent survey of faculty working with first-year graduate and undergraduate students indicated that, particularly at the graduate level, a problem still exists regarding significant numbers of students who gain entrance by TOEFL testing overseas, but do not meet the requirements of their departments. Liaison with these departments is underway to identify additional screening strategies which might be applied.

A study of grade point averages (GPA) of a random sample of undergraduate students (N = 117) in the Intermediate (21.150) and Advanced (21.190) EAP programme indicates a significant increase in overall GPA (p <.05) [Fox & Pychyl: 1991]. These students were placed by the CAEL, and their performance was tracked as they moved from the EAP programme into regular studies. The increase, from a "C" to a "B", took place in spite of an increase in the total number of courses allowed at the advanced level. This suggests some predictive validity. Other tracking studies (on a case study basis) are also underway. The results from these studies are not yet fully analyzed, but will be available early in 1993.

The CAEL provides a positive link between teacher designed activities in the EAP classroom, the university as a whole, and the student-learner who is working to meet the linguistic demands of academic study. Because EAP teachers are involved in test development, and because test development involves interaction with the university academy as a whole through active liaison between professors from all faculties on campus, the washback effects are positive and generative. This type of on-going collaboration certainly benefits the learners who are affected by CAEL decisions, but it is also of great benefit to professors -- both EAP professors and discipline-specific professors, who, through collaboration, develop a clearer understanding of how best to address ESL learners' needs with regard to academic study.

Because of the nature of CAEL, information about relative strengths and weaknesses in English can be provided to both the students taking the CAEL, and their professors. A profile like the one below is made available to CAEL testees so that CAEL can serve formative and diagnostic purposes. As a follow-up, counsellors are available to discuss CAEL results with students and to make suggestions regarding helpful learning strategies. [See Figure 5 below.]

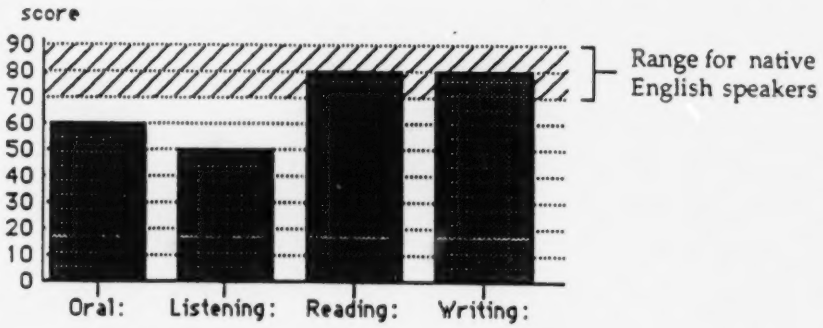
Figure 5

CALS Placement

Carleton University Assessment of English as a Second Language

Test Number: 011892-40

Oral: 60
Listening: 50
Reading: 80
Writing: 80



Placement:

A. Eligible to be considered for admission to full-time study in a degree program without any further ESL training.

ESL Courses Required:

- ☐ 21.196*
- ☐ 21.150
- ☐ 21.195
- ☐ 21.130
- ☐ 21.190
- ☐ Pre-Credit Intensive

Both Bachman [1990] and MessicK[1988] have identified logical, ethical and empirical considerations as sources of evidence to support a given interpretation or use of a test. As a langauge test, CAEL demonstrates ethical and logical aspects of validity. Unfortunately, a test that is derived logically from language in use creates interesting problems with regard to empirical analysis. This is particularly so for an analysis of the reliability of placement decisions.

A framework for discussing reliability and validity of CR test interpretations

Traditionally, that is, in norm-referenced weak true-score measurement models, when determining the quality -- "goodness" or "badness" -- of a test, reliability is considered a necessary condition for a good test. In fact, due to attenuated correlations and when empirical estimates of validity are evasive, reliability is considered a sufficient condition for a good test. However, reliability and validity cannot be considered as distinct concepts, but rather, they should be considered as two correlated dimensions within a space representing *inference* from test scores (i.e., test score interpretation). That is, the dimension of reliability tells the test constructor the relation between the test score and the score obtained under ideal conditions wherein the examinee was to answer the same set of items with the memory of the first test administration erased. On the other dimension, validity is concerned with the generalization of the test performance to other items, the entire domain of interest, or external criteria.

Zumbo[1] (1992), made the case for judging the quality of a test or measure by its meaningfulness and functional utility within the testing context. Zumbo's fundamental postulate is that the validity of a test or measure is not solely judged upon its fidelity with a latent structure and how well it preserves some unknown (and, in fact, unknowable) latent structure. Rather, the test constructor and scholarly community impinge meaning and functional utiility upon a test or measure. Meaningfulness and functional utility were broadly defined as:

(a) academic and/or social (societal) usefulness and importance of the inferences made from the test scores,

(b) substantial validity coefficients and indices (such as, concurrent, predictive, or factorial indices of validity),

(c) substantial and appropriate reliability or dependability of the measures,

(d) appropriate norms if a norm-referenced test, or

(e) appropriate cutoff-scores if its a criterion-referenced test,

---

[1]Written as a response to fundamental measurement's focus on levels of measurement and representational meaningfulness. This builds upon the work of Messick (1988).

(f) the ease (financial and otherwise) of administration and scoring of a test, and

(g) the appropriateness and necessity of the statistical measurement model used to justify knowledge claims about the test and inferences from the test.

All of these aspects of meaningfulness and functional utility are highly interconnected and represent a systemic understanding of meaningfulness. In this framework reliability speaks to the validity of a test in a different manner than it does in the traditional conceptions of validity. Reliability is now not a precursor of validity but an integral aspect of validity.

Reliability or dependability of the CAEL assessment

A. The Reliability or dependability of CR tests

Given our framework this section of the paper will focus on examining the reliability of the CAEL assessment as an integral part of amassing evidence for the validity of the assessment process. The reliability of a criterion-referenced (CR) test, such as the CAEL, is framed within the terms of the consistency of classification. Classification being the primary purpose of the CAEL assessment. That is, if the examinee was to answer the same set of items with the memory of the first test administration erased would they be classified within the same band or category as in the first administration. It should be noted that classical measures such as KR20 or Cronbach's Alpha are not meaningful within a CR approach. However, as in classical (weak) test theory, reliability of a CR test can be estimated with multiple-administrations of a test (or more precisely parallel forms) or with a single-administration of the test.

With the multiple-administration approach the researcher obtains usually two data sets which represent different forms of a test and/or times of administration. A comparison of the scores from the two sets of data assesses the generalizability over items and time. And in our context it will also address the issue of whether different versions of the test are equivalent.

For the single-administration approach the commonly known problems in multiple-administration -- time lag and intervention -- are avoided. However, although the number of individuals classified into each category is known, the joint distribution representing the

classifications is not. The most mathematically thorough solution to this problem has been given by Huynh (1976a; 1976b; 1977) with the use of the beta-binomial error model[2]. Another solution, not as thoroughly developed as the Huynh model, is given by Subkoviak (1980).

Just as there are many estimates of reliability (or dependability) in norm-referenced tests, there are many measures of reliability for CR tests. One of the classifiers of CR reliability indices is the loss function utilized with the index[3]. The loss functions commonly utilized in CR reliability are the threshold and squared-error loss functions. The threshold loss function was selected for our purposes because during score interpretation for the CAEL an individual score is referenced to a cut-off score, and that in using the CAEL one is interested in broad classifications -- mastery, nonmastery, and a mid-range classification -- not a degree of mastery or nonmastery along a continuum.

With the multi-administration approach to estimation of the reliability and the use of a threshold loss function there are two commonly used indices of reliability. The indices are the Hambleton & Novick (1973) index (p_o), which is an unbiased estimate of the reliability, measures the proportion of individuals consistently classified into the categories on repeated measurements, and the Swaminathan, Hambleton, & Algina (1974) index (K) based on Cohen's Kappa. The Swaminathan et al. index measures the proportion of individuals consistently classified over and above that expected by chance. The chance value is calculated as if the two random variables representing the repeated measurements were probabilistically independent.

The single-administration Huynh-approach with a threshold loss function allows one to calculate coefficients similar to p_o and K. That is, the beta-binomial error model is utilized to approximate the scores on the nonexistent second testing (as in a test-retest setting). Then, given the availability of the approximated second form, indices similar to those used in the multi-

---

[2]To acknowlege Huynh's contribution to the single-administration decision consistency problem the group of mathematical solutions to the single-administration problem will be referred to as the 'Huynh-approach'.

[3]Given that the domain score estimation (point estimation) is not a primary focus of the CAEL test the domain score estimation statistics based on G-theory were not considered. We concur with Hambleton & Novick (1973) who argued that the deviation of an examinees score from the cut-off score is not as important as whether or not the examinee is consistently assigned to the same side of the cut-off score on repeated testing.

administration approach can be calculated. To use the Huynh-approach one needs to have a test which is scored dichotomously and wherein the items are probabilistically independent.

B. Reliability of the CAEL assessment

Due to its format the CAEL test poses an interesting psychometric problem. The psychometric obstacles -- hurdles -- stem from the fact that the test was created with an eye toward remaining faithful to the role delineation (job analysis) rather than conforming to any given psychometric statistical model. That is, first, the CAEL test is comprised of tasks which are not scored dichotomously but rather with partial credit. Second, the tasks (items) in the CAEL test are not statistically independent rather they are clustered[4] or context dependent item sets (Haladyna, 1992) and therefore do not meet the assumption of local independence. Therefore, the single-administration models developed by Huynh are not directly applicable to the CAEL assessment. The multiple-administration indices are not effected by the item format because these analyses are not conducted at the item level and the joint distribution of scores is obtained via the retest.

The reliability results presented in this paper assume that the tasks are a representative sample selected from the domain defined by the role delineation (job analysis) described above. For all present purposes, it will be assumed that the cut-off scores described in Fox & Pychyl (1991) are appropriate for the context at Carleton University. Without this assumption any reliability measures from this data become uninterpretable. Furthermore, the term parallel version of a test should not be interpreted in a classical test theory sense where items are chosen to have equal means, standard deviations, and intercorrelations. Rather, criterion-referenced (CR) tests such as the CAEL are considered parallel if two or more tests are composed of randomly selected items (or as close as possible to random) from the same domain.

Two studies were conducted. Study one used extant data in a single-administration approach and study two utilized a test-retest strategy to assess the reliability of the CAEL assessment. It

---

[4] Some of the items require that the examinee get earlier items correct so that later items can be answered correctly. Therefore, the probability of getting later items correct require that you get other items correct. This is an example of the violation of independence. The job analysis indicated that this format reflects what students face in an academic setting.

should be recalled that reliability of the CAEL test refers to the decision consistency of classification.

Study One: Single-administration

This study was not of interest to us for the reliability results, per se. Rather, this study presents a first approximation to what we see as the very vexing problem of assessing the reliability of the CAEL test with one administration.

Method

Sample

Participants were 463 non-native speakers of English who had taken the CAEL in 1989 - 1990. We obtained data from ninety-two participants who had taken the Environment (also referred to as Rainforest) version of the CAEL. We also obtained data from 267 participants who took the Weather (also referred to as Elnino) version and 104 participants who had taken the Poplulaton version (also referred to as Overpopulation) .

Procedure

The tests were administered in the same format and in the same manner as a CAEL assessment would normally be administered. Each examinee completed the oral, writing, reading, and listening component of the CAEL assessment. However, the oral and writing components were not used for the single-administration analyses because they are comprised of essentially one task or item (i.e., an overall score is given for each component). The single-administration statistics were calculated for reading and listening components of the CAEL.

Reliability measures

The biggest obstacle in the single-administration is the lack of the joint distribution of classifications. Given that we are working with extant data we decided to examine split-half estimates of reliability. It should be recalled that if the split-half estimates are not stepped-up they

are severe lower-bound estimates of the reliability. We considered the following possble ways of dealing with the problem in the CAEL context (please recall that we have partial-credit scoring and clustered items):

1. Deriving a Huynh-type "beta-multinomial" model which would allow for the clustering of the items. However, the derivation resulted in an unmanageable expression which, in fact, could not be estimated. That is, it was computationally laborious and due to the item clustering the expression became mathematically intractable. The mathematical route was not fruitful.

2. We considered dichotomously coding the partial credit scheme so that a response to one item or task would have either {0,0} for incorrect response, {0,1} for half-credit, or {1,1} full credit. The test length would then be conceptually doubled and the binomial-error model could be applied. However, although at first blush this seems to solve the problem, the clustering of the items will not allow the beta-binomial model to be used. Also, this would give an inflated value of the reliability -- due to doubling the number of items -- and an appropriate step-down function is difficult to derive.

3. What appeared to be the most promising -- however, in the longrun it was not -- was a computer-intensive strategy wherein all possible split-halves were created and p_o and K would be calculated at each sub-sampling. Then an average would be taken over the samplings. Furthermore, one could calculate a standard error over all of the replications and use this measure in an assessment of the significance. There are

$$\frac{\frac{1}{2}K!}{\left[\left(\frac{1}{2}K\right)!\right]^2}$$

possible split-halves, where k represents the number of items in the test. Therefore, although computaionally laborious, this approach is feasible. The limitations are three-fold. First, due to the item clustering and the items vary in difficulty, not all splits-halves are sensical and therefore difficult to justify. Furthermore some Kappa values could be negative. Therefore some monitoring scheme would have to be built-in to allow only for "sensical split-halves". This is

computationally difficult to perform. Second, a step-up function is necessary. Hambleton (1992; personal comunication) clearly stated that an appropriate step-up function for the split-half is currently not available. Finally, the results would be conditional upon the examinees at hand. One would have to make certain that the examinees are a unbiased and representative sample of the population of examinees or generalization would be limited.

4. The most intuitively appealing and defensible approach was to conduct a split-half which is sensitive and aware of the test format and content. That is, the test constructor splits the test scale in half attempting to balance the total score of each scale and to create two conceptually equivalent halves. A four-by-four contingency table is obtained and the p_0 and K measures would be computed. However, given that these estimates are extreme lower bounds (given the halved number of items) some sort of step-up function needs to be imposed. Assuming that the items in each split-half are of the same kind a we can apply the simplistic Spearman-Brown-like formula

$$r''=\frac{2r}{1+r}$$

where r" is the estimated reliability of the full length test, and r is the split-half reliablity. The above step-up function is not optimal, however, it does give an estimate (or better stated, guess-timate) of the full-length reliability. To clearly indicate that function above is not optimal we have playfully labelled the step-up function the Crude and Barbaric step-up. If the two halves are not equivalent this function will overestimate the full-length reliability. The more conventional Spearman-Brown formula cannot be applied in this context because measures such as the correlation between the two halves are not meaningful in a CR context.

Considering all of the advantages and disadvantages, the method described in point 4 above was the one utilized in study one. The test constructors divided each subtest for each version of the CAEL into two equivalent halves paying particular attention to balancing for content and for total score. The cut-off scores (Fox & Pychyl, 1991) were corrected for reduced test length (i.e., halved) and the classfications were imposed on each for each subtest and for the overall placement.

The Hambleton & Novick index (p_o) and the Swaminathan, Hambleton, & Algina index (K) based on Cohen's Kappa were computed for each of the split-halves for each of the scales and each of the versions.

Results & Conclusions

For each of the listening and reading scores a four-by-four cross-classification matrix was computed. The results of the analyses are given in Table 1.

Insert Table 1 about here

In Table 1, p_o is the Hambleton & Novick consistency estimate, the crudely stepped-up p_o is the stepped-up value decribed in point 4 above, chance is the chance classification index, and Kappa is the Swaminathan et al. decision consistency index described above.

The results should be interpreted such that p_o is an extreme lower-bound and the stepped up value is a optimistic guess based on p_o. Chance should be interpreted as the correct classification if the two halves where independent and Kappa is a measure of consistency over and above chance. Kappa must be interpreted with the value of chance in consideration. Therefore, for the listening component of the Environment version the extreme lower-bound is .565, the optimistic consistency estimate is .722 and the Kappa is .315 over and above the chance classification of .365.

These results should be considered as a first approximation of the consistency of the CAEL assessment. They should not be interpreted as an accurate reflection of the reliability of the CAEL. The results are interesting only in demonstrating the application of methods in point 4 above. Given the problems in the single-administration assessment of the reliablity a two-administraton test-retest approach was used.

Study Two: Test-retest

Method

Sample

Participants were non-native speakers of English who had taken the CAEL within three weeks. They were told that they would have the opportunity to take the CAEL assessment for a second time. If they improved, their better score would stand. Eighty examinees took the Weather version of the CAEL assessment. Of the eighty, forty-five had taken the Environment version one week earlier and thirty-five had taken the Food (also referred to as Irradiation) version of CAEL three weeks before.

Procedure

The retest was administered in the same format and in the same manner as a CAEL assessment would normally be administered. Furthermore, the retest was scored without knowledge of who the examinee was. That is, the tests were scored and all data was coded by subject number. Each examinee completed the writing, reading, and listening component of the CAEL assessment on the retest. The oral component was not measured on the retest. Therefore, for the overall placement on the retest the oral score on the first administration was used. For the first and second administration each examinee was placed at one of the four categories discussed above.

Reliability measures

The Hambleton & Novick index (p_o) and the Swaminathan, Hambleton, & Algina index (K) based on Cohen's Kappa were computed for the entire sample and for each of the subsamples of 1-week (environment & weather versions) and 3-week (food & weather versions) time intervals from retest.

Results & Conclusions

For each of the writing, listening, reading, and overall placement scores a four-by-four cross-classification matrix was computed. Table 2 contains the Hambleton & Novick, and the Swaminathan, Hambleton, & Algina indices.

Insert Table 2 about here

In Table 2, p_o is the Hambleton & Novick consistency estimate, *chance* is the chance classification index, and *Kappa* is the Swaminathan et al. decision consistency index described above. The columns *increase* and *decrease* represent the proportion of individuals who move up in their classification or move down in their classification upon retesting. The results in Table 2 should be interpreted in the context of the sample sizes. That is, for small sample sizes a small shift in the cells will result in a dramatic change in the measure of proportion. As well, for the Kappa coefficient (for studies 1 and 2) the statistical significance was not calculated due to the unknown sampling structure. This is a conversative position of statistical significance testing.

The writing subtest is the key component of the CAEL for overall placement. The p_o index suggests that overall and for the two time intervals the consistency of classification is quite adequate. The 1-week and 3-week time intervals also suggest that the Environment and Weather versions are similar in classifying examinees, and that the Food and Weather versions are less similar than the Environment and Weather in their classifications. Examining the *increase* and *decrease* indices it appears that Food version is a bit more difficult than the Weather. That is, 17.2% of the examinees increased their classification from the Food to Weather versions. The p_o should be interpreted in conjunction with the Swaminathan et al. (Kappa) corrected index. The Kappa index, like the p_o, suggests that the Writing subtest is adequate in its classification consistency. Furthermore, the Kappa index also suggests that the Food and Weather versions are not as similar in their classifications as the Environment and Weather versions.

The Listening subtest shows an interesting phenomenon wherein the overall classification consistency is middling to moderate, however, when the sample is analysed according to equivalence of versions it is clear that the Food and Weather versions are less than moderate in their consistency of classification. Examining the *increase* and *decrease* indices it appears that Weather version is more difficult than the Food. That is, 45.7% of the examinees decreased in their classification from the Food to Weather versions. The Kappa index supports the Hambleton

& Novick index in that the Food and Weather versions are less than moderate in their decision consistency, however, the Environment and Weather are more than adequate.

The Reading subtest demonstrates moderate reliability overall and with the two time intervals (or interpreted as the equivalency of versions). The Kappa index also supports the moderate standing of the Reading subtest.

The overall placement is the most important measure for the CAEL assessment. Given that the overall placement is done via a profile over the four subtests, the subtests can fluctuate in their classification consistency but the overall placement must meet the most stringent of standards. Table 2, in fact, shows that the overall placement consistency is more than adequate overall and for the three different versions. An interesting benefit of the CAEL is demonstrated with the test-retest results. That is, focussing on the 3-week interval (the Food and Weather versions), the Food is more difficult than the Weather version for the writing subtest, however, for the listening subtest the opposite is true and the Weather version is more difficult than the Food version. However, overall placement is largest (with a 100% consistent classification) for the Food and Weather versions. The subcomponents can vary but the overall placement is the same. The profile aspect of the CAEL reflects a true component of language proficiency. That is, examinees who have the same overall score on a language test may differ greatly in their oral, writing, listening, and reading; however, they have developed skills in some areas of their language performance so as to compensate for weaknesses in others.

General Discussion

The purpose of this paper has been to present the foundation of the CAEL assessment and to present some preliminary findings as to its reliability of classification and hence aspects of its validity. The CAEL represents a natural coupling of ecologically rich (often termed authentic) tasks and criterion-referenced measurement.

The CAEL is founded on a thorough job analysis or in criterion-reference terms a role delineation of English language use at university. The development of the test reflects an

uncompromising belief that language testing should reflect language use. It is uncompromising because as Study 1 demonstrated, the test developers did not want to compromise support for validity, given by the job analysis, for the ease of computation and asssessment of reliability. The aspect of the CAEL that complicates the psychometrics is that it contains item sets which are context dependent, hence violating local independence, and partial credit scoring. The combination of no local independence and partial-credit scoring is a formidable hurdle for psychometric analysis. However, the job analysis indicated that this combination of item formats is commonly found in the university context.

The purpose of Study 1 is to explore some methods of single-administration reliability assessment. The first study is more interesting for the problems it posed rather than the results it produced. Study 2, however, demonstrated that the CAEL, although varying in its reliability of its subtests, has very good decision consistency overall -- in fact, well over 95% consistency. We do acknowledge that the sample size is not large enough and that replication is essential.

From a psychometric point of view, the CAEL assessment highlights an interesting foundational question. That is, when using statistical models to assess the psychometric properties of a test should the test format (i.e., format of the items, scoring, etc.) be modified so that the statistical models can be used or should the role delineation or job analysis have precedence over the statistical models? Given that we are striving for meaningful tests and ones which demonstrate functional utility (Zumbo, 1992), the role delineation has precedence over the statistical models. However, unlike some versions of authentic testing the psychometric properties of the CAEL are not discounted. One must assess the reliability of the decisions made with the test. Rather than seeing the nonconformity to psychometric models as a weakness of the test, we see it as a challenge (or call) to psychometricians to develop models which can be used to analyze tests like CAEL in order to determine its reliability and hence support validity and meaningfulness. Study 1 was but a first rough approximation to a solution. Further development is necessary in this area.

One aspect of the meaningfulness or functional utility (Zumbo, 1992) of the CAEL is its ability to consistently classify individuals on an overall basis. Note that this is not all of the functional

utility rather we must consider this within the light of the fact that the test is serving the academic community at Carleton University and that the university has continued its commitment both financially and personally. As a whole it is clear that the CAEL assessment has shown sound measurement properties. As well, the CAEL is an example of participatory model of the development of tests; hence, encouraging ownership by the professors and upper administration of the university.

However, further development is needed for the CAEL assessment. Clearly, a single-administration reliability method needs to be further explored. The two-administration approach is costly to both the test user and the examinee (recall that we were not able to retest on the oral component, the cost was very difficult to meet). As well, the procedures as being implemented so as to have subject/discipline specific tests. That is, for example, if an Engineering student wants to take the CAEL they will have the option to take an Engineering version of the CAEL. The examinee will then feel that they are taking a test which is representative of their area of interest. Finally, given that we began the reliability studies with the assumption that the cut-off scores are appropriate for Carleton, a standarding setting study and review is being planned for the next year. The development and validation of sound standards (cut-off scores) is essential to assessing the reliability the placement decisions.

References

Bachman, L. F. (1990) *Fundamental Considerations in Language Testing.* Oxford: Oxford University Press.

Britton, J. (1980). Shaping at the Point of Utterance, in Aviva Freedman & Ian Pringle (Eds.), *Reinventing the Rhetorical Tradition.* Conway, Arkansas: Council of Teachers of English: 61-66.

Fox, J., & Pychyl, T. (1991). *The Carleton Academic English Language Assessment: Linking Testing and Learning.* Paper presented at the 13th Annual Language Testing Research Colloquium at Educational Testing Service, Princeton, New Jersey.

Haladyna, T. M. Context dependent item sets. *Educational Measurement: Issues and Practice, 11,* 21-25.

Hambleton, R. K., & Novick, M.R. (1973). Toward an integration of theory and method for criterion-referenced tests. *Journal of Educational Measurement, 10,* 159-170.

Huynh, H. (1976a). On the reliability of decisions in domain-referenced testing. *Journal of Educational Measurement, 13,* 253-264.

Huynh, H. (1976b). Statistical consideration of mastery scores. *Psychometrika, 41,* 65-78.

Huynh, H. (1977). Two simple classes of mastery scores based on the beta binomial model. *Psychometrika, 42,* 601-608.

Messick, S.A. (1988). Validity. In R. L. Linn *Educational Measurement (3rd Edition)* New York: American Council of Education, Macmillan. (pp. 13-103).

Subkoviak, M. J. (1980). The reliability of mastery classification decisions. In R. A. Berk (Ed.), *Criterion-Referenced Measurement: The state of the art.* Baltimore: The Johns Hopkins University Press.

Swaminathan, H., Hambleton, R. K., & Algina, J. (1974). Reliability of criterion-referenced tests: A decision-theoretic formulation. *Journal of Educational Measurement, 11,* 263-267.

Zumbo, B. D. (1992). Meaningfulness and functional utility of tests. *Carleton Papers in Applied Language Studies.* This issue.

Table 1. Reliability estimates with extant data.

Test & scale	p_o	Crudely stepped-up p_o	Chance classification	Kappa
Listening				
Environment	.565	.722	.365	.315
Weather	.640	.781	.268	.509
Population	.558	.716	.455	.188
Reading				
Environment	.478	.647	.317	.236
Weather	.479	.648	.322	.233
Population	.654	.791	.502	.305

Table 2. Test-retest reliability results.

Subtest	subsample	p_o index	Chance	Kappa	Increase	Decrease
Writing[1]	total	.813	.441	.664	.112	.075
	1-week	.844	.445	.720	.067	.089
	3-week	.771	.438	.593	.172	.057
Listening	total	.638	.399	.397	.074	.288
	1-week	.800	.461	.629	.044	.156
	3-week	.429	.329	.148	.114	.457
Reading	total	.638	.453	.338	.174	.188
	1-week	.622	.470	.288	.20	.178
	3-week	.657	.438	.390	.143	.20
Overall Placement	total	.988	.433	.978	.012	.000
	1-week	.978	.409	.962	.022	.000
	3-week	1.000	.462	1.000	.000	.000

1. Writing is the key subcomponent of the CAEL for overall placement. As well, difficulty varies for some subtests, however, this reflects the reality in the academy.